HITIQA: Towards Analytical Question Answering

Sharon Small[1], **Tomek Strzalkowski**[1], **Ting Liu**[1], **Sean Ryan**[1], **Robert Salkin**[1], **Nobuyuki Shimizu**[1], **Paul Kantor**[2], **Diane Kelly**[2], **Robert Rittman**[2], **Nina Wacholder**[2]

[1]The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
{small,tomek,tl7612,seanryan,
rs6021,ns3202}@albany.edu

[2]Rutgers University
4 Huntington Street
New Brunswick, NJ 08904
{kantor,diane,hitiqa,
wacholder}@scils.rutgers.edu

Abstract

In this paper we describe the analytic question answering system HITIQA (High-Quality Interactive Question Answering) which has been developed over the last 2 years as an advanced research tool for information analysts. HITIQA is an interactive open-domain question answering technology designed to allow analysts to pose complex exploratory questions in natural language and obtain relevant information units to prepare their briefing reports. The system uses novel data-driven semantics to conduct a clarification dialogue with the user that explores the scope and the context of the desired answer space. The system has undergone extensive hands-on evaluations by a group of intelligence analysts. This evaluation validated the overall approach in HITIQA but also exposed limitations of the current prototype.

1 Introduction

Our objective in HITIQA is to allow the user to submit exploratory, analytical questions, such as "What has been Russia's reaction to the U.S. bombing of Kosovo?" The distinguishing property of such questions is that one cannot generally anticipate what might constitute the answer. While certain types of things may be expected (e.g., diplomatic statements), the answer is heavily conditioned by what information is in fact available on the topic. From a practical viewpoint, analytical questions are often underspecified, thus casting a broad net on a space of possible answers. Questions posed by professional analysts are aimed to probe the available data along certain dimensions. The results of these probes determine follow up questions, if necessary. Furthermore, at any stage clarifications may be needed to adjust the scope and intent of each question. Figure 1 shows a fragment of an analytical session with HITIQA; note that these questions are *not* aimed at factoids, despite their simple form.

User: *What is the history of the nuclear arms program linking Iraq and other countries in the region?*
HITIQA: [responses and clarifications]
User: *Who financed the nuclear arms program in Iraq?*
HITIQA*:…*
User: *Has Iraq been able to import uranium?*
HITIQA:*…*
User: *What type of debt does exist between Iraq and her trading partners in the region?*

FIGURE 1: A fragment of an analyst's session with HITIQA

HITIQA project is part of the ARDA AQUAINT program that aims to make significant advances in the state of the art of automated question answering. In this paper we focus on three aspects of our work:

1. Question Semantics: how the system "understands" user requests
2. Human-Computer Dialogue: how the user and the system negotiate this understanding
3. User Evaluations and Results

2 Factoid vs. Analytical QA

There are significant differences between factoid, or fact-finding, and analytical question answering. A *factoid question* is normally understood to seek a piece of information that would make a corresponding statement true (i.e., it becomes a fact): "How many states are in the U.S.?" / "There are X states in the U.S." In this sense, a factoid question usually has just one correct answer that can generally be judged for its truthfulness with respect to some information source.

As noted by Harabagiu et al. (1999), factoid questions display a distinctive "answer type", which is the type of the information item needed for the answer, e.g., "person" or "country", etc. Most existing factoid QA systems deduct this expected answer type from the form of the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
HITIQA: Towards Analytical Question Answering	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The State University of New York at Albany,1400 Washington Avenue,Albany,NY,12222**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**7**	

question using a finite list of possible answer types. For example, "Who was the first man in space" expects a "person" as the answer type. This is generally a very good strategy that has been exploited successfully in a number of automated QA systems, especially in the context of TREC QA[1] evaluations. Given the excellent results posted by the best systems and an adequate performance attained even by some entry-level system, we believe that the process of factoid question answering is now fairly well understood (Harabagiu et al., 2002; Hovy et al., 2000; Prager at al., 2001, Wu et al., 2003).

In contrast to a factoid question, an analytical question has a virtually unlimited variety of syntactic forms with only a loose connection between their syntax and the expected answer. Given the many possible forms of analytical questions, it would be counter-productive to restrict them to a predefined number of question/answer types. Therefore, the formation of an answer in analytical QA should instead be guided by the user's intended interest expressed in the question, as well as through any follow up dialogue with the system. This clearly involves user's intentions (the speech acts) and how they evolve with respect to the overall information strategy they are pursuing.

In this paper we argue that the semantics (though not necessarily the intent) of an analytical question is more likely to be deduced from the information that is considered relevant to the question than through a detailed analysis of its particular form. We noted that the questions analysts ask, while clearly part of a strategy, are generally quite flexible and "forgiving", in the sense that there is always a strong possibility that the answer may not arrive in the expected form, and thus a change of strategy, and even the initial expectations, may be warranted. This suggests strongly that a solution to analytic QA must involve a dialogue that combines information seeking and problem solving strategies.

3 Document Retrieval

HITIQA works with unstructured text data, which means that a document retrieval step is required to detect any information that may be relevant to the user question. It has to be noted that determining "relevant" information is not the same as finding an answer; indeed we can use relatively simple information retrieval methods (keyword matching, etc.) to obtain perhaps 200 "relevant" documents from a database. This gives us an initial information space to work on in order to determine the scope and complexity of the answer, but we are nowhere near the answer yet. The current version of HITIQA uses the INQUERY system (Callan et al., 1992), although we have also used SMART (Buckley, 1985) and other IR systems (such as Google).

4 Text Framing

In HITIQA we use a text framing technique to delineate the gap between the possible meaning of the user's question and the system "understanding" of this question. We can approximate the meaning of the question by extracting references to known concepts in it, including named entities. The information retrieved from the database may well lead to other interpretations of the question, and we need to determine which of these are "correct".

The framing process imposes a partial structure on the text passages that allows the system to systematically compare different passages against each other and against the question. Framing is not attempting to capture the entire meaning of the passage; it needs to be just sufficient enough to communicate with the user about the differences in their question and the returned text. In particular, the framing process may uncover topics or aspects within the answer space which the user has not explicitly asked for, and thus may be unaware of their existence. If these topics or aspects align closely with the user's question, (i.e., matching many of the salient attributes) we may want to make the user aware of them and let him/her decide if they should be included in the answer.

Frames are built from the retrieved data, after clustering it into several topical groups. Passages are clustered using a combination of hierarchical clustering and n-bin classification (Hardy et al., 2002a). Each cluster represents a topic theme within the retrieved set: usually an alternative or complimentary interpretation of the user's question. Since clusters are built out of small text passages, we initially associate a frame with each passage that serves as a seed of a cluster. We subsequently merge passages and their associated frames to arrive at one or more combined frames for the cluster.

HITIQA starts text framing by building a *general frame* on the seed passages of the clusters and any of the top *N* (currently *N=10*) scored passages that are not already in a cluster. The general frame represents an event or a relation involving any number of entities, which make up the frame's attributes, such as LOCATION, PERSON, ORGANIZATION, DATE, etc. Attributes are extracted from text passages by BBN's Identifinder, which

[1] TREC QA is the annual Question Answering evaluation sponsored by the U.S. National Institute of Standards and Technology www.trec.nist.gov

tags 24 types of named entities. The event/relation itself could be pretty much anything, e.g., *accident, pollution, trade*, etc. and it is captured into the TOPIC attribute from the central verb or noun phrase of the passage. In the general frame, attributes have no assigned roles; they are loosely grouped around the TOPIC (Figure 2).

We have also defined three slightly more specialized *typed frames* by assigning *roles* to selected attributes in the general frame. These three "specialized" frames are: (1) a *Transfer frame* with three roles including FROM, TO and OBJECT; (2) a two-role *Relation frame* with AGENT and OBJECT roles; and (3) an one-role *Property frame*. These typed frames represent certain generic events/relationships, which then map into more specific event types in each domain. Other frame types may be defined if needed, but we do not anticipate there will be more than a handful all together.[2] For example, another 3-role frame may be *State-Change frame* with AGENT, OBJECT and INSTRUMENT roles, etc.[3]

FRAME TYPE: *General*
TOPIC: *imported*
LOCATION: *Iraq, France, Israel*
ORGANIZATION: *IAEA* [missed: *Nukem*]
PERSON: *Leonard Spector*
WEAPON: *uranium, nuclear bomb*
DATES: *1981, 30 November 1990*, ..

FIGURE 2: A general frame obtained from the text passage in Figure 3 (not all attributes shown).

Where the general frame is little more than just a "bag of attributes", the typed frames capture some internal structure of an event, but only to the extent required to enable an efficient dialogue with the user. Typed frames are "triggered" by appearance of specific words in text, for example the word *export* may trigger a Transfer frame. A single text passage may invoke one or more typed frames, or none at all. When no typed frame is invoked, the general frame is used as default. If a typed frame is invoked, HITIQA will attempt to identify the roles, e.g. FROM, TO, OBJECT, etc. This is done by mapping general frame attributes selected from text onto the typed attributes in the frames. In any given domain, e.g., weapon non-proliferation, both the trigger words and the role identification rules can be specialized from a training corpus of typical documents and questions. For example, the role-id rules rely both on syntactic cues and the expected entity types, which are domain adaptable.

Domain adaptation is desirable for obtaining more focused dialogue, but it is not necessary for HITIQA to work. We used both setups under different conditions: the generic frames were used with TREC document collection to measure impact of IR precision on QA accuracy (Small et al., 2004). The domain-adapted frames were used for sessions with intelligence analysts working with the WMD Domain (see below). Currently, the adaptation process includes manual tuning followed by corpus bootstrapping using an unsupervised learning method (Strzalkowski & Wang, 1996). We generally rely on BBN's Identifinder for extraction of basic entities, and use bootstrapping to define additional entity types as well as to assign roles to attributes.

The version of HITIQA reported here and used by analysts during the evaluation has been adapted to the Weapons of Mass Destruction Non-Proliferation domain (WMD domain, henceforth). Figure 3 contains an example passage from this data set. In the WMD domain, the typed frames were mapped onto *WMDTransfer* 3-role frame, and two 2-role frames *WMDTreaty* and *WMDDevelop*. Adapting the frames to the WMD domain required very minimal modification, such as adding the WEAPON entity to augment the Identifinder entity set, generating a list of international weapon control treaties, etc.

The Bush Administration claimed that Iraq was within one year of producing a nuclear bomb. On 30 November 1990... Leonard Spector said that Iraq possesses 200 tons of natural uranium imported and smuggled from several countries. Iraq possesses a few working centrifuges and the blueprints to build them. Iraq imported centrifuge materials from Nukem of the FRG and from other sources. One decade ago, Iraq imported 27 pounds of weapons-grade uranium from France, for Osirak nuclear research center. In 1981, Israel destroyed the Osirak nuclear reactor. In November 1990, the IAEA inspected Iraq and found all material accounted for....

FIGURE 3: A text passage from the WMD domain data

HITIQA frames define top-down constraints on how to interpret a given text passage, which is quite different from MUC[4] template filling task

[2] Scalability is certainly an outstanding issue here, and we are working on effective frame acquisition methods, which is outside of the scope of this paper. While classifications such as (Levin, 1993) or FrameNet (Fillmore, 2001) are relevant, we are currently aiming at a less detailed system.

[3] A more detailed discussion of possible frame types is beyond the scope of the current paper.

[4] MUC, the Message Understanding Conference, funded by DARPA, involved the evaluation of information extraction systems applied to a common task.

(Humphreys et al., 1998). What we're trying to do here is to "fit" a frame over a text passage. This also means that multiple frames can be associated with a text passage, or to be exact, with a cluster of passages. Since most of the passages that undergo the framing process are part of some cluster of very similar passages, the added redundancy helps to reinforce the most salient features for extraction. This makes the framing process potentially less error-prone than MUC-style template filling.

A very similar framing process is applied to the user's question, resulting in one or more *Goal frames,* which are subsequently compared to the data frames obtained from retrieved text passages. A Goal frame can be a general frame or any of the typed frames. Goal frames generated from the question, *"Has Iraq been able to import uranium*?" are shown in Figures 4 and 5.

FRAME TYPE: *General*
TOPIC: *import*
WEAPON: *uranium*
LOCATION: *Iraq*

FIGURE 4: A general goal frame from the Iraq question

The frame in Figure 4 is simply a General frame which is invoked first. HITIQA then discovers that TOPIC=*import* denotes a Transfer-event in the WMD domain, so it creates a *WMDTransfer* frame that replaces the general frame. This new frame, shown in Figure 5, has three role attributes TRF_TO, TRF_FROM and TRF_OBJECT, plus the relation type (TRF_TYPE). Each role attribute is defined over an underlying general frame attribute (given in parentheses), which are used to compare frames of different types. The role-id rules rely both on syntactic cues and the expected entity types, which are domain adaptable.

FRAME TYPE: *WMDTransfer*
TRF_TYPE (TOPIC): *import*
TRF_TO (LOCATION): *Iraq*
TRF_FROM (LOCATION, ORGANIZATION): *?*
TRF_OBJECT (WEAPON): *uranium*

FIGURE 5: A typed goal frame from the Iraq question

HITIQA automatically judges a particular data frame as relevant, and subsequently the corresponding segment of text as relevant, by comparison to the Goal frame. The data frames are scored based on the number of conflicts found with the Goal frame. The conflicts are mismatches on values of corresponding attributes, specifically when the data frame attribute list does not contain any of the entities in the corresponding Goal Frame attribute list. If a data frame is found to have no conflicts, it is given the highest relevance rank, and a conflict score of zero.

All other data frames are scored with an increasing value based on the number of conflicts, score 1 for frames with one conflict with the Goal frame, score 2 for two conflicts etc. Frames that conflict with all information found in the query are given the score 99 indicating the lowest rank. Currently, frames with a conflict score 99 are excluded from further processing as outliers. The frame in Figure 6 is scored as relevant to the user's query and included in the answer space.

FRAME TYPE: *WMDTransfer*
TRF_TYPE (TOPIC): *imported*
TRF_TO (LOCATION): *Iraq*
TRF_FROM (LOCATION): *France*
TRF_OBJECT (WEAPON): *uranium*
CONFLICT SCORE*: 0*

FIGURE 6: A typed frame obtained from the text passage in Figure 3, in response to the Iraq question

5 Enabling Dialogue with the User

Framed information allows HITIQA to automatically judge text passages as fully or partially relevant and to conduct a meaningful dialogue with the user about their content. The purpose of the dialogue is to help the user navigate the answer space and to negotiate more precisely what information he or she is seeking. The main principle here is that the dialogue is primarily content oriented. Thus, it is okay to ask the user whether information about the AIDS conference in Cape Town should be included in the answer to a question about combating AIDS in Africa. However, the user should never be asked if a particular keyword is useful or not, or if a document is relevant or not.

Our approach to dialogue in HITIQA is modeled to some degree upon the mixed-initiative dialogue management adopted in the AMITIES project (Hardy et al., 2002b). The main advantage of the AMITIES model is its reliance on data-driven semantics which allows for spontaneous and mixed initiative dialogue to occur. By contrast, the major approaches to implementation of dialogue systems to date rely on systems of functional transitions that make the resulting system much less flexible. In the grammar-based approach, which is prevalent in commercial systems, such as in various telephony products, as well as in practically oriented research prototypes (e.g., DARPA Communicator; Seneff and Polifoni, 2000; Ferguson and Allen, 1998), a complete dialogue transition graph is designed to guide the conversation and predict user responses, which is

suitable for closed domains only. In the statistical variation of this approach, a transition graph is derived from a large body of annotated conversations (e.g., Walker, 2000; Litman and Pan, 2002). This latter approach is facilitated through a dialogue annotation process, e.g., using Dialogue Act Markup in Several Layers (DAMSL) (Allen and Core, 1997), which is a system of functional dialogue acts.

Nonetheless, an efficient, spontaneous dialogue cannot be designed on a purely functional layer. Therefore, here we are primarily interested in the semantic layer, that is, the information exchange and information building effects of a conversation. In order to properly understand a dialogue, both semantic and functional layers need to be considered. In this paper we are concentrating exclusively on the semantic layer.

6 Clarification Dialogue

The clarification dialogue is when the user and the system negotiate the information task that needs to be performed. Data frames with a conflict score of 0 form the initial kernel answer space and HITIQA proceeds by generating an answer from this space. Depending upon the presence of other frames outside of this set, the system may initiate a dialogue with the user. When the Goal frame is a general frame HITIQA first initiates a clarification dialogue on existing general data frames that have one conflict. All of these 1-conflict general frames are first grouped on their common conflict attribute. HITIQA begins asking the user questions on these near-miss frame groups, with the largest group first. The groups must be at least groups of size N, where N is a user controlled setting. This setting restricts of all HITIQA's generated dialogue. HITIQA then check for the existence of any data frames that are one of the three typed frames. Clarification dialogue will be initiated on these, when all of their general attributes agree with the general attributes of the Goal frame respectively. Alternatively, if the Goal frame is one of the three type specific frames, a clarification dialogue is first initiated on groups of one conflict data frames that are the same type as the Goal frame. The clarification dialogue will then continue to the remaining two type specific frames if any exist, and finally on to any General data frames.

A *1-conflict* frame has only a single attribute mismatch with the Goal frame. This could be a mismatch on any of the general frame attributes, for example, LOCATION, ORGANIZATION, TIME, etc., or in one of the role-assigned attributes, TO, FROM, OBJECT, etc. A special case arises when the conflict occurs on the TOPIC attribute, which indicated the event type. Since all other attributes match, we may be looking at potentially different events (though similar type events) involving the same entities, occurring at the same location or time. The purpose of the clarification dialogue in this case is to probe which of these additional events may be of interest to the user.

Another special case arises when the Goal frame has a different value of the TYPE attribute than a data frame. The purpose of the clarification dialogue here is to see if the user wishes to expand the answer space to include events of a different type. This situation is illustrated in the exchange shown in Figure 7.

The clarification question asked by HITIQA in Figure 7 was obtained by comparing the Goal frame (Figure 5) to a partly matching frame (Figure 8) generated from another text passage. We note first that the Goal frame for this example is *WMDTransfer* type, while the data frame in Figure 8 is *WMDDevelop* type. Nonetheless, both frames match on their general-frame attributes WEAPON and LOCATION. Therefore, HITIQA asks the user if it should expand the answer space to include development of uranium in Iraq as well.

User: *"Has Iraq been able to import uranium?"*

HITIQA: IRAQ REPORTED TO HAVE IMPORTED MISSILES AND URANIUM

Supporting passage:

...that **Iraq** possesses 200 tons of natural **uranium imported** and smuggled from several countries. **Iraq** possesses a few working centrifuges and the blueprints to build them. **Iraq** imported centrifuge materials from Nukem of the FRG and from other sources. One decade ago, **Iraq imported** 27 pounds of weapons-grade **uranium** from France, for Osirak nuclear research center...

HITIQA: *"Are you also interested in background information on the uranium development program in Iraq?"*

User: …

FIGURE 7: The clarification dialogue detail

During the dialogue, as new information is obtained from the user, the Goal frame is updated and the scores of all the data frames are reevaluated. If the user responds the equivalent of "yes" to the system clarification question in the dialogue in Figure 7, a corresponding *WMDDevelop* frame will be added to the set of active Goal frames and all *WMDDevelop* frames obtained from text passages will be re-scored for possible inclusion in the answer.

FRAME TYPE: *WMDDevelop*
DEV_TYPE (TOPIC): *development*, *produced*
DEV_OBJ (WEAPON): *nuc. weapons*, *uranium*
DEV_AGENT (LOCATION): *Iraq*, *Tuwaitha*
CONFLICT SCORE: *2*
Conflicts with FRAME_TYPE *and* TOPIC

FIGURE 8: A 2-conflict frame against the Iraq/uranium question that generated the dialogue in Figure 7.

The user may end the dialogue at any point using the generated answer given the current state of the frames. Currently, the answer is simply composed of text passages from the zero conflict frames. In addition, HITIQA will generate a "headline" for the text passages in the answer space. This is done using a combination of text templates and simple grammar rules applied to the attributes of the passage frame. Figure 7 shows a portion of the answer generated by HITIQA for the Iraq query.

7 HITIQA Preliminary Evaluations

We have evaluated HITIQA in a series of workshops with professional analysts in order to obtain an in-depth and comprehensive assessment of the system usability and performance. In addition to evaluating our research progress, the purpose of these workshops was to test several evaluation instruments to see if they can be meaningfully applied to a complex information system such as HITIQA.

For the participating analysts, the primary activity at these workshops involved preparation of reports in response to "scenarios" – complex questions that often encompass multiple sub-questions, aspects and hypotheses. For example, in one scenario, analysts were asked ti locate information about the al Qaeda terorist group: its membership, sources of funding and activities. In another scenario, the analysts were requested to find information on the chemical weapon Sarin. Figure 9 shows one of the analytical scenarios used in these workshops. We prepared a database of over 1GByte of text documents; it included articles from the Center for Non-proliferation (CNS) data collected for the AQUAINT program and similar data retrieved from the web using Google. The analysts' task was to prepare a report "as much like what you would do in your normal work environment as possible." Over the six days of the workshops, each analyst prepared five such reports in sessions of one to three hours. Each session involved multiple questions posed to the system, as well as clarification dialogue, visual browsing and report construction. Figure 10 shows an abridged transcript from another analytical session with HITIQA.

The department chief has requested a report by the close of business today on the nuclear arms program in Iraq and how it was influenced by the neighboring countries. List the extent of the nuclear program in each involved country including funding, capabilities, quantity, etc. Your report should also include key figures in Iraq nuclear program as well as in other countries in the region, and,any travels that these key figures have made to other countries in regards to a nuclear program, any weapons that have been used in the past by either country, any purchases or trades that have been made relevant to weapons of mass destruction (possibly oil trade, etc.), any ingredients and chemicals that have been used, any potential weapons that could be under development, countries that are involved or have close ties to Iraq or her trade partners, possible locations of development sites, and possible companies or organizations that these countries work with for their nuclear arms program. Add any other information relating to the Iraqi Nuclear Arms Programs.

Figure 9: A scenario level analytic task

One of our primary concerns was to design tasks that were similar in scope and difficulty to those that the analysts are used to performing at work and to ensure that they felt comfortable using the system. 5 questions in the scenario evaluation dealt with this issue; for example, one question asked how the scenarios compared in difficulty with the tasks the analysts normally perform at work. The mean score for these five questions was 3.75 on a 5 point scale (five is the best score). The lowest score (M=2.88) was received on the question 'How did the scenario compare in difficulty to tasks that you normally perform at work?"; this slightly above average rating of difficulty of the tasks was quite satisfactory for our purposes.

In the final evaluation, analysts were asked to rate their agreement with statements such as "Having HITIQA helps me find important information" (score 4.50), "Having Hitiqa at work would help me find information faster than I can currently find it" (score 4.33), and "Hitiqa would be a useful addition to the tools that I already have at work" (score 4.25). The mean normalized score for the combined final evaluation of Workshop I was 3.75 on the 5 point scale; this means that the system received many more ratings of 4 and 5 than of 1 and 2. Comments made by the analysts in the group discussion and in the individual interviews confirmed that analysts liked the interactive dialogue and were very pleased with the results. For example, one analyst said "I learned more about Sarin gas in 30 minutes than I probably would have at work in a half a day." As desired, the analysts also made many suggestions for improving the interface and the interoperation of

the visual and text display. For a research system undergoing its first rigorous evaluation, these results are very satisfactory – they support the value of the design of the HITIQA system, including the interactive mode and the visual display and encourage us to move forward with this approach.

User: *What is the status of South Africa's chemical, biological, and nuclear programs?*
Clarification Dialogue: 1 minute
Studying Answer Panel: 60 minutes
Copying 24 passages to report
Visual Panel Browsing: 5 minutes
User: *Has South Africa provided CBW material or assistance to any other countries?*
Clarification Dialogue: 1 minute
Studying Answer Panel: 26 minutes
Copying 6 passages to report
Visual Panel browsing: 1 minute
Adding 1 passage to report
User: *How was South Africa's CBW program financed?*
Clarification Dialogue: 40 seconds
Studying Answer Panel: 11 minutes
Copying 3 passages to report

FIGURE 10: Fragment of an analytical session

8 Future work

The AQUAINT Program has entered its second phase in May 2004. Over the next 2 years our focus will be on augmenting HITIQA to provide more advanced dialogue capabilities, including problem solving dialogue related to hypothesis formation and verification. This implies building up system's knowledge acquisition capabilities by exploiting diverse data sources, including structured databases and the internet.

9 Acknowledgements

This paper is based on work supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program. Special thanks to Heather McCallum-Bayliss and John Rogers for helping to arrange the analyst workshops. Additional thanks for Google for extending their license for this experiment, to Ralph Weischedel of BBN/Verizon for the use of IdentiFinder, to Chuck Messenger and Peter LaMonica for assistance in development of the analytical scenarios, and to Bruce Croft at University of Massachusetts for the use of INQUERY system.

References

Allen, J. and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. www.cs.rochester.edu/research/cisd.

Buckley, C. 1985. *Implementation of the Smart information retrieval system.* TR85-686, Computer Science, Cornell University.

Ferguson, G. and J. Allen. 1998. *TRIPS: An Intelligent Integrated Problem-Solving Assistant*. AAAI-98 Conf., pp. 567-573.

Fillmore, C. & C. F. Baker. 2001. Frame semantics for text understanding. *WordNet Workshop* at NAACL.

Hardy, H., et al. 2002a. *Cross-Document Summarization by Concept Classification*. Proceedings of SIGIR, Tampere, Finland.

Hardy, H., et al. 2002b. *Multi-layer Dialogue Annotation for Automated Multilingual Customer Service*. ISLE Workshop, UK.

Harabagiu, S., et. al. 2002. *Answering Complex, List and Context questions with LCC's Question Answering Server.* TREC-10.

Hovy, E., et al. 2000. *Question Answering in Webclopedia. Notebook*. Proceedings of Text Retrieval Conference TREC-9.

Humphreys, R. et al. 1998. Description of the LaSIE-II System as Used for MUC-7. Proc. of 7th Message Under. Conf. (MUC-7.).

Levin, B. 1993. *English Verb Class and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Litman, Diane J. and Shimei Pan. 2002. *Designing and Evaluating an Adaptive Spoken Dialogue System*. User Modeling and User-Adapted Interaction. Vol. 12, No. 2/3, pp. 111-137.

Prager, J. et al, 2003. *In Question-Answering Two Heads are Better Than One.* Proceedings of HLT-NAACL 2003, pp 24-31.

Seneff, S. and J. Polifroni. 2000. *Dialogue Management in the MERCURY Flight Reservation System*. ANLP-NAACL 2000.

Small et al. 2004. A Data Driven Approach to Interactive Question Answering. In M. Maybury (ed). *Future Directions in Automated Question Answering*. MIT Press (to appear).

Strzalkowski, T and J. Wang. 1996. A self-learning Universal Concept Spotter. Proceedings of COLING-96, pp. 931-936.

Walker, M. A. 2002. *An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email.* Journal of AI Research, vol 12., pp. 387-416.

Wu, M. et al. 2003. *Question Answering by Pattern Matching, Web-Proofing, Semantic Form Proofing*. TREC-12.Notebook.